DOCUMENT RESUME

ED 166 212 TM 007 812

AUTHOR Granville, Arthur C.; And Others
TITLE The Impact of Evaluation: Lessons Drawn from the Evaluations of Five Early Childhood Education Programs.
PUB DATE Mar 78
NOTE 45p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Decision Making; Early Childhood Education; Evaluation Methods; *Federal Programs; Information Dissemination; *Information Utilization; Performance Factors; *Policy Formation; *Preschool Programs; *Program Evaluation

ABSTRACT

Five different program evaluations were described to indicate those qualities which make an evaluation effective or not effective. Evaluation effectiveness was defined as impact on decision making or long-term policy formation, and influence upon a variety of audiences. Robert D. Matz described the First Chance Project, and concluded that the evaluation methodology used to inform policymakers should be distinct from the approach used to improve teaching. John M. Love, who was associated with the national Home Start Demonstration Program, felt that several factors contributed to the value and use of evaluation information: evaluations planned with the program; timely reports; rigorous experimental design; respect between agency and evaluators; demonstrated relationships between process and outcomes; and non-controversial appeal of the program. Project Developmental Continuity was discussed by Arthur C. Granville. Factors affecting evaluation impact included the relevance of quantitative data; sociopolitical acceptability of the implications; and pertinence to evaluation criteria. Allen G. Smith, who discussed Project Follow Through, supported close relationships between evaluation research and curriculum. Lawrence J. Schweinhart of the Ypsilanti (Michigan) Perry Preschool Project recommended intensive one-site studies, determination of the feasibility of longitudinal studies, and adequate funding. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED166212

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

High/Scope Educational Research Foundation
600 North River Street
Ypsilanti, Michigan 48197

THE IMPACT OF EVALUATION: LESSONS DRAWN FROM THE EVALUATIONS OF FIVE EARLY CHILDHOOD EDUCATION PROGRAMS

PROCEEDINGS OF A SYMPOSIUM PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, TORONTO, CANADA, MARCH 1978

Participants

Arthur C. Granville
John M. Love
Robert D. Matz
Lawrence J. Schweinhart
Allen G. Smith

High/Scope Educational Research Foundation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John M. Love

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Chairperson

Senta A. Raizen

National Institute of Education

TM007 812

Table of Contents

	Page
INTRODUCTION	1
Introductory Comments by Senta A. Raizen, Chairperson.	1
The Impact of the Evaluation of the First Chance Project, High/Scope Model *Robert D. Matz*	7
The Impact of the Evaluation of the National Home Start Demonstration Program *John M. Love*	15
The Impact of the Project Developmental Continuity Evaluation *Arthur C. Granville*	25
The Impact of the Evaluation of Project Follow Through Cognitively Oriented Curriculum *Allen G. Smith*	33
The Impact of the Evaluation of the Ypsilanti Perry Preschool Project *Lawrence J. Schweinhart*	41

INTRODUCTION

Evaluation is a growth industry. A recent review of governmental evaluation practice by Rein & White[1] suggests that it is now a matter of routine to write requirements for evaluation into any human welfare legislation. This growing emphasis is reflected in the fact that federal support for evaluations of non-defense programs increased by 500%--to $146,000,000--from 1969 to 1974[2]. Yet, as fashionable as it has become, evaluation has often made a poor showing when evaluated itself:

> *At the moment, there seems to be no evidence that evaluation, although the law of the land, contributes anything to educational practice other than headaches for the researcher, threats for the innovators, and depressing articles for journals devoted to evaluation.*[3]

> *The recent literature is unanimous in announcing the general failure of evaluation to affect decision making in a significant way.*[4]

> *We might as well be candid: federal evaluations so far have been largely ineffective.*[5]

What makes evaluation ineffective? What does it take to make it effective? This symposium offers answers derived from the participants' collective experience with five early childhood education programs whose evaluations were intended to have some measure of national impact.

---

[1]Rein, M. & White, S. H. Policy research: Belief and doubt. Policy Analysis, Spring 1977, 239-271.

[2]Kelezo, G. J. Program evaluation: Emerging issues of possible legislative concern relating to the conduct and use of evaluation in the Congress and Executive Branch. Washington, D.C.: Congressional Research Service, Library of Congress, 1974.

[3]Rippey, R. M. What is transactional evaluation? In R. M. Rippey (Ed.) Studies in transactional evaluation. Berkeley: McCutchan, 1973.

[4]Wholey, J. S., Scanlon, J. W., Duffy, H. G., Fukimoto, J. S., & Vogt, L. M. Federal evaluation policy. Washington, D.C.: The Urban Institute, 1973.

[5]Bruce, R. G. What goes wrong with evaluations and how to prevent it. Human Needs, 1972, 1.

Nature of the presentations. All of the symposium participants are staff members of the High/Scope Educational Research Foundation, under whose auspices these evaluations have been conducted. However, each participant has been involved principally with one of the evaluations and represents that one in this symposium. Thus while the participants have encountered many common issues, they have done so independently, and their reactions to these encounters are diverse. The participants and the evaluations (funding sources indicated parenthetically) whose impact they discuss are:

- Robert D. Matz, The Impact of the Evaluation of the First Chance Project, High/Scope Model (Bureau of Education for the Handicapped);
- John M. Love, The Impact of the Evaluation of the National Home Start Demonstration Program (Administration for Children, Youth, and Families);
- Arthur C. Granville, The Impact of the Project Developmental Continuity Evaluation (Administration for Children, Youth, and Families);
- Allen G. Smith, The Impact of the Evaluation of Project Follow Through, Cognitively Oriented Curriculum (U. S. Office of Education);
- Lawrence J. Schweinhart, The Impact of the Evaluation of the Ypsilanti Perry Preschool Project (Carnegie Corporation).

Objectives of the symposium. The five investigators, examining the respective programs with which they have been principally associated, (a) assess the impact that the evaluations have had upon their audiences, (b) analyze the reasons for the impact, or lack of it, and (c) draw implications for the evaluation of early childhood education programs in particular and educational programs in general.

Readers are reminded that the papers presented here are not evaluation reports. It would be more accurate to describe them as evaluations of evaluation reports. Those who are interested in the original program evaluations that these presentations reflect on are invited to write to the authors for the relevant documents.

Introductory Comments by Senta A. Raizen, Chairperson
National Institute of Education
Washington, D.C.

I am very pleased to be able to chair this session, which has been organized by a group of researchers from the High/Scope Foundation. Each of the participants has been responsible for a particular evaluation of some of the development work of High/Scope or has led a project assessing an independent early childhood education program. The reason for presenting these papers as a group, and they will be given in very abbreviated form so that you can hear all of them, is to take a hard look at the impact of evaluation--of a set of evaluations. We all know that there has been increased questioning about the value of evaluation, what it is that evaluation can contribute to decision making, whether it affects further research and development, whether it influences policy. And despite the skepticism, the increasing skepticism, there is an increasing demand by Congress, the federal government, and others for accountability, for assessment, for knowing the "value" returned from funded programs. More and more money is being invested in evaluation studies at the same time that we are questioning their results and effects. Hence the interest and topicality of this set of papers which speculate about the issue of impact, effect and value of evaluation.

As we listen to these papers, I would like to call your attention to some themes that they have in common. Therefore, I would like the audience to hold their discussion questions until all the papers have been presented.

First, pay attention to what the authors have to say, or not say, about the audiences they thought the evaluation was addressing. Was the originally intended audience the same as the audiences eventually addressed? Did the evaluation end up not addressing that audience but having value to another, perhaps unanticipated? For any evaluation, possible audiences might be:

- Federal program officials, as in the case of these evaluations, almost all of which dealt with federally funded programs;
- Local implementors of innovations or of new curricula, local school people, people who run preschools, etc.;
- Researchers and developers in the general field--in early childhood in this case, but in whatever pertinent component of the R&D community;

- Peers relating very closely to the evaluation--in this case, the High/Scope Foundation fellow researchers;

- Powerful "lay" audiences who are neither educators, education program officials, nor education researchers; (An example in one paper was the Office of Management and Budget that was to be convinced of the efficacy of particular program.)

- General lay audiences--people who in the end elect officials to Congress, to state legislatures, to local school boards, and so forth.

A second theme that runs through the papers is a concern with the substance of a particular evaluation: What were the intended learnings? What were the objectives of the evaluation? What type of evaluation was it? Did the author start out with one kind of objective and end up with another? Did he always know at the outset, or did he only conclude in retrospect, what objectives the evaluation served? Some alternative objectives represented by this group of evaluations are inherent in the following questions:

- Were there any decisions to be made that were to be influenced by the evaluation?

- Was the evaluation to capture details of the program so one could understand more clearly its nature?

- Was it to judge the "goodness" of the program (summative evaluation)?

- Was it aimed at program improvement (formative evaluation)?

- Was it supposed to affect R&D, allow others to develop better curricula, different types of tests, evaluation designs, what not?

- Was it to have an effect on national or local policies?

- Was it to demonstrate that we know what to do in early childhood education to make it effective?

- Was the evaluation supposed to be used for "do or don't", "increase or decrease" kinds of program decisions?

A word about methodology: I don't know how many of you had a chance to read Ernie House's recent article in the Educational Researcher.[1] He has developed a typology, or summarized typologies developed by others, of different types of evaluation methodologies. So pay a little attention to the type of methodology used in each study: At one end of the continuum,

---

[1]House, E. R. Assumptions underlying evaluation models. Educational Researcher, March 1978, 4-12.

we have rigorous experimental design, using random samples, pre/post tests, hard outcome measures. At the other end, we have methodologies derived from other fields: art criticism, the adversary mode borrowed from the court room, transactional analysis. In what way was methodology related to the purpose of an evaluation? Did the methodology have an influence on who used evaluation results? Do you agree with the author's hunches, which I found very interesting, as to how process affected impact?

A last theme I would like to raise with the audience as well as the authors is the following: How realistic are our expectations about the likely effects of evaluations, and, in fact, of research in general, of which evaluation is a special instance? Too often, our expectations are for immediate use of a particular evaluation or a particular piece of research. Recently, some researchers (Rich at Princeton and Caplan at Michigan) have made a distinction between the instrumental use of knowledge and the conceptual use of knowledge. The former entails the immediate application of information, and the immediate visibility of a consequence for a program, i.e., a program is changed, a project is abolished, an activity is increased--something happens because of that specific evaluation, that piece of research. Alternatively, in the conceptual use of knowledge, (including knowledge that comes out of evaluations) one doesn't see anything happening as a direct effect of a piece of work on the program on which it reports or which it purports to be influencing. But some years later, one observes some paradigm shifts, concepts have changed, some attitudes have shifted, and people really are doing things differently in the field. That is surely an important effect of research and of evaluation, one we tend to neglect when we look for impact.

A second point related to our exaggerated notions of potential imapct has to do with our understanding of decision-making processes. Though groups of people playing different roles may vary in how they reach decisions, there are always many more factors that go into decision making than just systematically derived and reported information and knowledge. There are political processes, consensual processes, that allow for an accomodation of different views and interests. The methodologies that we use as social scientists are only a part--and should be only a part--of decision making and policy formulation. We need to remember our limited role.

A third issue concerns time frames: Perhaps we look too soon for the effects of a particular piece of research. One of the authors makes the point well, and I'll leave it to him. A last concern with respect to our expectations that comes out of rural sociology research is the social distance question: How far in social distance, in point of reference, is the evaluator from the audience he is trying to address? Are the authors here saying, "Gee, we had a lot of effect on the High/Scope researchers, but we didn't have very much effect on OMB?" Are we being

unrealistic when we expect effects across what are really rather large social distances? Should we not be satisfied if we can have an effect on our peers? And if that is not enough, how can we translate/interpret our findings so as to speak to audiences removed from us? And if we cannot, are there intermediaries who could?

With that, let me introduce the authors. They will give their papers in the order listed in your program.

The Impact of the Evaluation of the First Chance Project[1,2,3] High/Scope Model

Robert C. Matz
High/Scope Educational Research Foundation
Ypsilanti, Michigan

The First Chance Project is intended to develop effective strategies for working with handicapped children during the preschool years. Each of the individual projects has primary responsibility for demonstrating effectiveness.

The High/Scope First Chance Project

The High/Scope BEH-funded First Chance Project supported a demonstration preschool classroom from 1974-1977 with two half-day sessions five days per week. Approximately 15 children, one-third of whom were moderately to severely handicapped, attended each session with a team of two teachers which taught only one session per day. In terms of social-economic status, the sample of children was heterogeneous, being divided about equally between middle- and lower-income children. Since the era of the Ypsilanti Perry Preschool Project, which began in 1962, the High/Scope curriculum framework (known as the Cognitively Oriented Preschool Curriculum) has been evolving. The curriculum framework supports universal developmental processes by creating an environment in which teachers and children initiate active learning experiences.

Audiences for the Evaluation

The major goal of the project was to demonstrate the effectiveness of the curriculum as a framework for serving an integrated group of handicapped and non-handicapped children. Two major types of audience were considered--the ultimate targets of the curriculum, teachers and curriculum developers, and governmental policymakers, particularly the Joint Dissemination Review Panel (JDRP) of the U.S. Office of Education and the National Institute of Education.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

[2]The work summarized here was supported by Grant OEG-0-74-2720 from the United States Office of Education, Bureau of Education for the Handicapped.

[3]Thanks to James T. Bond for providing many of the ideas in this paper and providing illustrative examples.

From the outset, we have seen approval ("validation") by JDRP as a critical step in building the future of the program. Attaining this validation would promote further support of our demonstration preschool including the development of instruments, media, teacher training materials, and so forth. Additionally, JDRP approval would increase the likelihood of other preschool centers adopting the Cognitively Oriented Curriculum. Thus, we have accepted literally the JDRP guidelines. According to these, programs must demonstrate growth in children's skills/abilities that is attributable to the process of education.

The Evaluation

The evaluation was multi-faceted, including instrument development, assessment of teaching competencies, and fine-grained naturalistic observation of social interaction as well as standardized testing. Since the evaluation has just recently been completed, in most cases judgments of its effects would be premature. Accordingly this paper describes the ways in which child progress was evaluated, reports the results, and examines closely one area in which some conclusions can be drawn regarding impact. This is the effects of the evaluation upon teachers.

Our principal means of addressing the JDRP guidelines was collecting pre- and post-test data with the McCarthy Scales of Children's Abilities, a norm-referenced test with five separate scales measuring verbal, quantitative, perceptual-performance, memory, and motor abilities. The use of norms provided a way around the immense difficulty of finding a comparison group with a blend of characteristics comparable to those of the children in the program, and the apparent potential for diagnosis afforded by the separate scale scores allowed us to assess the wisdom of using the results for formative purposes. In a different context, the McCarthy scales provided a benchmark for examining the validity of new instruments.

Analyses of changes in scale scores showed that important gains were accomplished by a large proportion of children. As we expected, the largest gains occurred during children's first year in the program. Middle-income children registered as much improvement as lower-income children. This fact is of interest since most evaluations of preschool programs have been concerned with lower-income populations. The majority of handicapped children accomplished gains as well. And given the nature of the handicapping conditions of those children who did not accomplish gains, it may be unrealistic to expect any preschool program to accelerate development in the areas measured by the McCarthy Scales.

With the intention of informing a current "burning issue" in early childhood education, we examined the social integration of handicapped and non-handicapped children. A checklist of social and emotional behaviors was developed and used in conjunction with naturalistic observations for two years in two classrooms each year. During both years, most of the behaviors observed indicated that handicapped children were well integrated with their classmates and that teachers generally responded in the same ways to both groups.

New child-outcome measures were developed during the course of the evaluation in response to our belief that existing measures do not encompass the breadth of the teaching/learning process. The narrowness of existing measures precludes summative assessment in vital curriculum areas and limits formative interchange between teaching practice and evaluation results.

The Child Observation Record (COR) is a set of observation categories and rating scales completed by teachers for each child at the beginning, middle, and end of the school year. For the majority of COR items, the teacher indicates the child's highest level of functioning that occurs on at least three days of a five-day observation interval. COR items tap these abilities:

- ability to structure own activity;
- ability to create non-verbal representations;
- sophistication of dramatic play;
- ability to define and solve problems;
- ability to control attention;
- interest in the processes of reading and writing;
- ability to engage in cooperative play;
- affective quality of peer relations;
- pride in own work;
- intrinsic motivation;
- social problem-solving skills.

Although the summative use of the COR depends upon a comparison group or the collection of baseline data, in our case the fact that COR results generally parallel the McCarthy Scale results allows us to utilize COR data in documenting claims of program impact.

Work in devising the Preschool Productive Language Assessment Task (PPLAT) was begun to meet the widely felt need for relatively brief but valid assessment of young children's communicative competence. Cross-project cooperation with Follow Through and Project Developmental Continuity staff facilitated greatly this work. Our major conclusions regarding the PPLAT were in the area of concurrent validity. We found that the McCarthy Scale General Cognitive raw scores predict almost 50% of the variance in PPLAT scores and that two communication-related COR items predict an additional 20% of tha variance. The hallmark of the PPLAT is a relatively open measurement situation in which the child has an opportunity to initiate interactions with the tester and with real objects. Whether or not such a situation will permit the reliability demanded of summative procedures is unclear.

Effects of the Evaluation

The two categories of audience referred to in this paper, policymakers on one hand and on the other, teachers and curriculum developers, are not exhaustive and are really continuous in some respects. For instance governmental officials do not comprise the only influence upon policy; the field of early childhood education, itself, including teachers, has certain leverage. But preserving the duality will simplify discussion of the effects of the evaluation. These audiences have overlapping but basically different needs regarding the substance of evaluation. For the former audience, evaluation must provide diagnostic/formative input for daily planning and long-range modifications of entire instructional systems. For policymakers, evaluation must assess capacities that are valued by a political constituency and that are or could be influenced by educational experience.

Impact Upon Teachers

A distinctive element of our evaluation enterprise was the emphasis we placed upon meeting the needs of teachers and fitting the processes of assessment with classroom processes. Achieving this fit is a different matter with the High/Scope curriculum (and others resembling it) than with either traditional or programmed curricula. In the Cognitively Oriented Preschool Curriculum, the aim of teaching is to build upon, not simply respond to or direct, the child's actions, ensuring that they become a context for thinking. Toward this end, teaching teams engage in extensive planning and evaluation sessions on a daily basis keeping organized, albeit informal, written records. Teachers' planning and evaluation is guided by a set of developmentally-based "key experiences" through which children's abilities can be broadened and consolidated and by a group of teaching strategies focusing upon principles of room arrangement, daily routine, and so forth.

Working within this framework, we have needed to ask ourselves, *"What does it mean for a formal evaluation to have an impact upon teachers?"* We agree with the notion that evaluation feedback should be integrated in the instructional process. Ideally formative and summative evaluation should form two sides of the same coin. But we are discomforted with the constraining effect that standardized tests have upon teaching.

Tests constrain teaching. When feedback from formal assessment procedures is integrated with instructional process, teaching tends to become geared to the same highly controlled situations as those involved in the assessment procedure. That is, classroom process begins to resemble alternate forms of whatever test is used. This brings with it three distinct problems:

- There is an increase in the amount of teacher-directed activity at the expense of child-directed and mutually initiated activity.

- Successful performance may indicate mastery of a trivial rote procedure rather than underlying knowledge.

- The attention of teachers, peers, and adults may be directed to a child's weaknesses rather than his strengths.

And these problems are as likely to accompany the use of criterion- and objective-referenced as norm-referenced tests.

With this in mind, we decided against conveying to teachers McCarthy Scale results for individual children. We believed that input from teachers' daily planning and evaluation (and occasional consultations with experts providing supplementary services to children such as physical therapy) would be sufficient for devising appropriate teaching strategies. And we conducted an evaluation substudy to probe this assumption.

Our research hypothesis was that children whose assessment data might have been given to teachers would show improvement in their apparent deficit areas without the actual delivery of feedback. That is, the children in question were those whose MSCA pretest performance revealed severe developmental incongruities, which we defined as having scale scores <-2 standard deviations (SD) for one or two of the five individual scales. According to current schools of thought, knowledge of the MSCA results for this group of children would be especially valuable for teachers, permitting them to concentrate upon the areas in which special assistance is needed. The pretest performance of eight children belonged in the category of interest. Of these children, only two had been classified as handicapped upon program entry. In all cases, save one, the children did raise their scores above -2 SD on the posttest. (The exception was a child with a serious heart problem who was developmentally delayed on the Motor scale). Thus, we found no evidence that the diagnostic information provided by the McCarthy Scales is necessary for improving children's performance in areas where they had particularly low pretest scores. Given the widely recognized drawbacks of labeling children in terms of their test scores (e.g., Hobbs, 1977), the implication is that the evaluation had a positive effect upon teachers by <u>not</u> doing something, by not relaying the McCarthy Scales data.

<u>Usefulness of the COR</u>. In contrast, by devising the COR we actively influenced the teaching process. In curricula resembling the High/Scope model, systematic teacher observations have direct bearing upon the organization of classroom experience, while the information provided by most, if not all, one-shot tests is distant and antagonistic to recommended teaching strategies. The COR is intended to focus teachers' attention upon general categories of behavior, thereby guarding against the dangers of teaching to the test. Collection of COR data was encouraged by teacher trainers and was found by teachers to have a number of valuable uses:

- Completing the COR helps the teacher know the child's interests and abilities and plan for the child's needs;
- Completing the COR gives training in curriculum areas;
- Completing the COR helps the teacher perceive child progress;
- COR information provides a springboard for communication with parents about the child's actions in the preschool and at home.

Sections of the COR have been included in a Michigan State Department of Education manual on child observation techniques.

Impact Upon Policymakers

A paper for JDRP is being prepared currently, so the jury still is out regarding the JDRP validation. The presentation of evaluation results in publications (e.g., High/Scope, 1977; Ispa & Matz, 1977) and at the annual High/Scope Spring Conference has helped to acquaint state and district officials with the power of the curriculum. Supporting this assertion is the fact that these decisionmakers have either commenced or continued to work with us in disseminating the curriculum model. The PPLAT also gives promise of having impact upon policymakers. The Administration for Children, Youth, and Families (ACYF) is considering the PPLAT for use in the national evaluation of a bilingual preschool demonstration project.

Implications

A lesson learned from this evaluation is that it might be best to separate the evaluation methodology used for informing policymakers from the process of teaching. In other words, assuming our definitions of the needs of both audience types, it seems impossible to satisfy these needs with a unitary approach. And this will remain true unless alternative approaches to summative evaluation, such as the PPLAT, are validated by careful research and accepted by educational authorities. Achieving a summative evaluation system that has relevance for teacher planning in cognitively oriented curricula will require a fundamental reshaping of the concepts underlying current approaches. Instead of a situation in which curricula shrink to fit tests, instrumentation must expand to encompass the richness of the curricula.

In comparison with large-scale evaluations of educational impact, the High/Scope First Chance Project had a high ratio of senior research staff to subjects. Additionally, these staff were interested in child development, had previous experience teaching preschoolers, and were willing to work closely with teachers and children. Reciprocally, the teachers involved recognized

a need for the evaluation work that was undertaken. This permitted us to engage in a wide number of applied activities interrelated by two integrating questions: *How can we devise evaluation procedures more in concert with the aims of the curriculum than those now available* and *how can we learn about the relative effectiveness of alternative methodologies?* Coincidentally, this evaluation paradigm is consistent with what Weisberg and Bryk (1978) perceive as a much needed and emerging trend in special education research. After a duration of five to ten years, we should be able to determine whether this general approach actually has enriched educational practice.

References

Hobbs, N. The futures of children: Categories, labels, and their consequences. San Francisco: Jossey-Bass, 1975.

High/Scope Educational Research Foundation. Demonstration preschool project final program performance report. Ypsilanti, Mich.: Author, 1977.

Ispa, J., & Matz, R. D. Integrating handicapped preschool children within a cognitively oriented program. In M. J. Guralnick (Ed.), Early intervention and the integration of handicapped and non-handicapped children. Baltimore, Md.: University Park Press, in press.

Weisberg, H., & Bryk, J. Unpublished discussion paper. Cambridge, Mass.: Huron Institute, 1978.

The Impact of the Evaluation of the National Home Start Demonstration Program[1,2]

John M. Love
High/Scope Educational Research Foundation
Ypsilanti, Michigan

The Home Start Demonstration Program was begun by the Office of Child Development (now the Administration for Children, Youth, and Families) in 1971 to demonstrate "alternative ways of providing Head Start-type comprehensive services for young children in their homes." Concurrent with program planning, a comprehensive evaluation was also planned, an evaluation that would involve process, cost and summative data. The demonstration phase of the program ended in spring 1975, the first report of program impact appeared in October 1974, and the final report of the evaluation was published in March 1976, giving us a reasonable time interval for assessing the impact of the evaluation.

Description of the Home Start Program

Sixteen projects geographically distributed across the United States were funded for a three-year period to serve about 80 families each. Eligibility guidelines were the same as for Head Start; the average Home Start family had four or five members who subsisted on incomes of less than $6,000 per year. All children in the family between the ages of 3 and 5 were considered "focal" children and received full program benefits; children under 3 were also included to some extent. The typical Home Start project consisted of a director, three specialists (a nurse, social service/parent involvement coordinator and either a home visitor supervisor or an education specialist), seven home visitors, and a secretary. Each home visitor provided services to 10 or 11 families.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

[2]The work summarized here was performed pursuant to Contract No. HEW-105-72-1100 with the Office of Child Development, Department of Health, Education and Welfare by the High/Scope Foundation and Abt Associates Inc. The findings only briefly described here are summarized in Love, J. M., Nauta, M. J., Coelen, C. G., Hewett, K., & Ruopp, R. R. National Home Start Evaluation Final Report: Findings and Implications. Ypsilanti, Mich.: High/Scope Foundation and Abt Associates Inc., March 1976.

Home Start projects were shaped by federal guidelines that specified activities and services in four areas (education, health, nutrition and social/psychological services)[3]. Program services in all areas were provided through regular home visits, group activities for parents and children, and through arrangements with community agencies. During home visits, the home visitor focused activities primarily on the parent. Our observations showed that about one third of the interactions were between home visitor and parent, one third between home visitor and child, and one third involved all three participants. Although over half the activities related to educational concerns, most home visitors followed a curriculum that insured attention to all four program components. Most projects supplemented home visits with monthly group activities for parents and children, a particularly important process for the families located in isolated communities or rural areas.

Overview of the Evaluation

The evaluation was designed to measure changes in parents and children (the summative evaluation), to determine the relative cost-effectiveness of Home Start and Head Start, and to assess program process. In addition to its descriptive function the process evaluation permitted examination of relationships between programmatic variables, parent and child outcomes, and costs. Six sites were selected for inclusion in the summative evaluation since there were not sufficient resources for including all 16, and although the site selection process was by no means random, analyses of project characteristics showed no major differences between the six summative sites and the other ten.

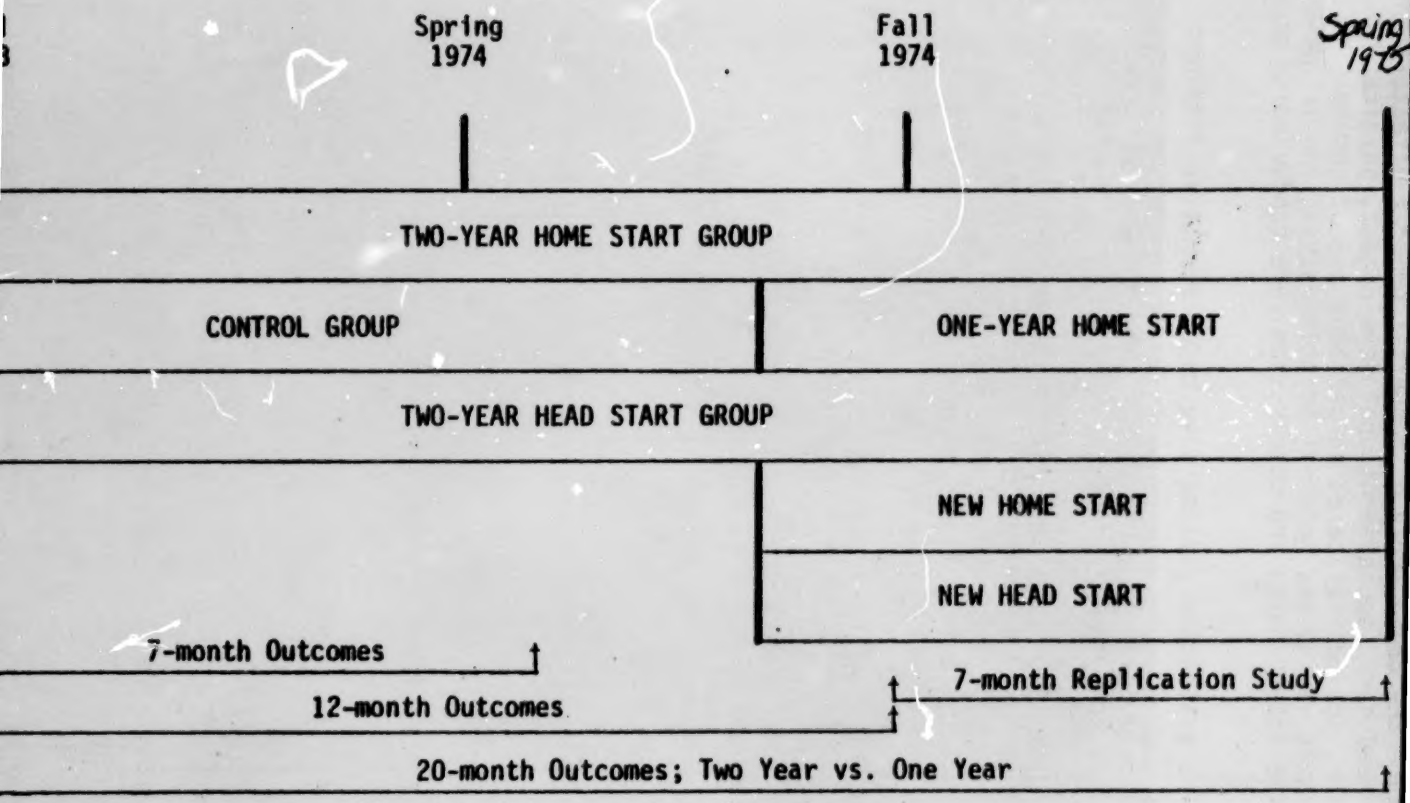
The evaluation was designed to compare the Home Start group with a no-treatment control group and with a Head Start group. Pretesting was conducted in fall 1973, with posttests in spring 1974 (referred to in our reports as the seven-month outcomes) and fall 1974 (12-month outcomes). The control group was selected through an over-recruitment procedure whereby each project located twice as many families as could be served, and a random half were asked to wait one year before entering the program. Although there were practical considerations that prevented strict adherence to the randomization procedures, later analyses revealed virtually no differences between the Home Start and control groups in their entering characteristics. When the control group entered Home Start their first year data provided an estimate of one-year program effects for comparison with the two-year (1973-75) experience of the original Home Start group (see Figure 1). Families were not randomly assigned between Home Start and Head Start, so the comparison of those two groups was complicated by some initial sample differences (e.g., Home Start families were more disadvantaged than Head Start families).

---

[3]Officially, parent involvement is included as a Home Start/Head Start component, but we always considered parent involvement to permeate the entire Home Start program, rather than to be a separate component.

Figure 1

SUMMATIVE EVALUATION DESIGN AND ANALYSES



rows indicate testing times; horizontal lines joining arrows indicate intervals between
ints, labeled with the analysis that was performed.

The summative evaluation used eleven distinct measures in an attempt to capture the broad array of potential program effects--on children's school readiness, social-emotional development, physical development, nutrition, and medical care; on parent's use of community resources, home environment variables, maternal teaching style, and mother-child relationship. Extensive interviewing and observation were conducted to obtain the process data.

The evaluation yielded five major summative findings:

- Home Start produced significant changes in parents (compared to the control group) after both 7 and 12 months;
- Home Start children realized significantly greater gains than control children, after both 7 and 12 months, in such areas as school readiness, social-emotional ratings, and medical and dental care, although the 7-month findings were stronger than the 12-month results;
- One year of participation in Home Start appeared to be about as beneficial as two years;
- Home Start and Head Start had virtually equivalent effects on parents; and
- Home Start and Head Start children showed virtually equivalent gains.

A large number of programmatic findings from the process evaluation were also reported. Among the more important of these were the following:

- Home visit focus on parents increased over the course of the program (in relation to focus on the child);
- The number of specialists employed (an important cost factor) in various service delivery areas varied considerably across projects;
- Projects were successful in obtaining considerable resources from community sources (roughly 20% of the total program resources).

Of particular interest were findings relating various elements of program process and showing relationships between process and outcomes; for example:

- Home visitors working with more than 13 families had difficulty maintaining frequent and regular contact with families;

- Amount of home visit time spent on a particular component was not associated with parent or child outcomes in that area (e.g., home visitors spending consistently more time in school readiness activities did not have children who scored higher than home visitors who spent less time in this area);

- Variations in frequency and duration of home visits did have an effect on parent and child outcomes (e.g., both frequency of visits and average duration of visits were significantly related to some outcomes).

Consequences of the Evaluation

With some degree of hindsight it has been possible to list some of the effects the Home Start evaluation has had. The effects seem to fall into four categories--the evaluation affected (1) the demonstration program itself, in a formative fashion, (2) other Head Start or Head Start-related programs, (3) the development of evaluation instruments, and (4) the implementation of evaluation designs for policy research. The effects in each of these areas are discussed next, followed by an attempt to understand why this particular evaluation has apparently had so many positive consequences.

Formative Effect on the Demonstration Program Itself

Although it may seem natural for a formative evaluation to serve a formative function, in large-scale, multi-site projects there are many obstacles to this ideal arrangement. In the case of Home Start, however, an attentive program officer in Washington read all evaluation reports and took immediate action where it seemed warranted. Two examples provide a flavor of the evaluation's impact. In the first year of the study, data from the Food Intake Questionnaire (a 24-hour recall method of assessing children's diets) indicated that children were not receiving the daily recommended amounts in the basic food groups; the federal program officer sent directives to all programs alerting them to this finding and recommending that more attention be given to nutrition component activities.

As part of the cost study, we were collecting information on how program staff spent their time (in order to allocate costs by program components). These data showed that supervisors of home visitors did not often accompany home visitors on their visits, but instead did most of their "supervision" in the office by reviewing records, discussing visits with the home visitors, and so on. Upon seeing this finding, the federal program officer requested that the Home Start programs spend more time on in-the-home supervision.

Head Start and Head Start-Related Programs

Although a basic purpose of the evaluation was to obtain information that would help guide decisions about a home-based effort within Head Start, we did not expect decisions to be made before the evaluation was completed. But even during the course of the evaluation, many local Head Start programs began adding a home-based component to their operation, or at least some home-based activities for some of the Head Start families. This, of course, could not have been due to any evaluation findings, yet the very existence of a major demonstration project with its accompanying evaluation provided considerable publicity for the approach of working with families in their homes; people apparently liked the idea, and did not require data in order to act.

Two very concrete effects of the completed evaluation can be documented--(1) programmatic findings relating to the number of families home visitors should serve, the frequency of visits and the amount of supervision staff should receive, were written into official guidelines for Head Start programs operating home-based options; (2) when ACYF established regional training centers to train staff for home-based programs, evaluation findings were used to set standards for training sessions (it was required that every session include a review of evaluation findings and that every trainee receive a copy of the evaluation's final report). The findings must have been considered useful and relevant by officials in the Administration for Children, Youth, and Families if they found their way into official program guidelines and requirements.

One of the concrete products of the formative evaluation was a book presenting the views and experiences gathered from Home Start staff. The book turned out to be more than a documentation of the Home Start experience. It has been printed by the Government Printing Office and is available as a resource to anyone interested in the operation of home-based programs.[4]

Impact on Instrumentation for Similar Evaluations

This area may be viewed as only marginally critical when it comes to assessing the contribution of a particular evaluation. Nevertheless, related evaluations (such as all evaluations of Head Start, or all Title I studies) are frequently criticized for lack of uniformity in their dependent measures. Thus it is at least worth noting that subsequent studies have, in some cases, used instruments first developed for Home Start, and in other cases, decided not to use instruments because of the evidence from Home Start (although the latter decisions are more difficult to document).

---

[4]Hewett, K. D., Jerome, C. M., Nauta, M., Rubin, A. D., & Stein, M. Partners with Parents: The Home Start experience with preschoolers and their families. U.S. Government Printing Office, 1978. (DHEW Publication No. DHEW (OHDS) 78-31106.)

Two measures that were developed for Home Start to address outcome areas often inadequately treated in large-scale evaluations have since found wider use. A rating scale of social-emotional development, the High/Scope Pupil Observation Checklist (POCL), has now been used in at least three other ACYF studies--the evaluations of two demonstration programs (Project Developmental Continuity, by High/Scope, and the Child and Family Resource Program, by Development Associates) and the major National Day Care Center Study (by Abt Associates). The High/Scope Home Environment Scale has been used by Applied Management Sciences in an ACYF study of mainstreaming in Head Start.

Implementation of Designs for Policy Research

It is difficult to ascertain the extent to which other evaluation designs have been influenced, but to the extent the Home Start evaluation provides a "successful" example of the application of a true experimental design to answer questions about social programs, it may have influenced others. In designing an evaluation of bilingual curriculum projects ACYF is currently planning to use a true experiment. In addition, Home Start provided a successful example of using case studies for assessing program process.

Speculation on Reasons for These Consequences

It is probably not possible to be very certain as to why these effects have occurred. There are a number of features of the evaluation and its context, however, that suggest some plausible reasons.

Experimental Design

First, the classic textbook experimental design was applied and, to all intents and purposes, implemented; the procedures appeared to have been cleanly executed and people felt good about it.

Process-Outcome Findings

Second, the design permitted conclusions about relationships between process and outcome. Without this, guidelines could not have specified the number of families that should be served, etc.

Advance Planning

Third, the evaluation was designed simultaneously with the design of the demonstration project. Program and evaluation staff within the federal agency worked together at the planning stage, thus avoiding a large number of problems that plague other studies. There are at least two possible implications of this collaboration: the program staff understood the importance of working within certain design requirements (such as the delayed-entry control group), and evaluation staff better understood what type of findings would be most relevant.

Unambiguous Findings

Fourth, the findings showed clear-cut treatment effects. It seems to me less likely that this study would have influenced others if the program itself had not been found effective.

Program Appeal

Related to this is the fact that people found the program appealing in and of itself. The idea of working directly with families, rather than through a center-based operation, struck a responsive chord in large numbers of practitioners. The zeitgeist undoubtedly set the stage for accepting evaluation findings.

Nature and Timing of Reports

The effects of the evaluation may have also been due to the nature and timing of reports. Thus, a sixth reason may be that evaluation reports were submitted in draft form within six months of the completion of data collection; the relevant information was available, along with the tables and charts to add appropriate status, at a time when it was still useful to those managing the programs. Seventh, the reports were generally considered to be fairly readable. The final report in particular was presented as a relatively brief, non-technical document that could be widely circulated. In fact, there is evidence that it is being read by persons working in home-based programs.

Conclusion

In reviewing the four years of collaboration between ACYF officials and evaluation contractor staff, I can't escape the conclusion that the impact of the Home Start study has had a lot to do with the quality of interactions that existed between the two parties and the mutual respect

that developed. Without federal officials who really responded to findings and made serious efforts to act on them when it seemed prudent, evaluation staff may have become less "involved" in efforts to make the evaluation relevant. Without the evaluation contractors' efforts to understand the program and make their work useful, federal officials might have been more inclined to dismiss uncomplimentary findings. The responsibility for making evaluations effective seems to fall on both sides, and in the case of Home Start we were fortunate to experience this mutual concern that, I believe, paid off in the consequences described in this paper. If one can legitimately translate these experiences into prescriptions for program evaluations, we would conclude with the following. Evaluation findings are most likely to be of value and actually used:

- When planned in conjunction with the program;
- When relationships between process and outcomes are revealed;
- When the program itself has wide appeal or is non-controversial;
- When reports are timely, relevant and readable;
- When a rigorous experimental design is implemented; and
- When there is mutual respect between agency program officials and evaluation contractors.

The Impact of the Project Developmental Continuity Evaluation[1,2]

Arthur C. Granville
High/Scope Educational Research Foundation
Ypsilanti, Michigan

Description of the Program

Project Developmental Continuity (PDC) is an experimental program designed to strengthen connections between Head Start centers and public schools. The ultimate purpose of strengthening these connections is to help ensure continuity of experience for Head Start children as they progress upward through school. And continuity of experience, in turn, will enhance the children's social competence, or at least this is the hope that motivated the Office of Child Development to develop and fund the program in thirteen sites around the country, beginning in 1974. (The Office of Child Development has since been renamed the Administration for Children, Youth, and Families--ACYF.) The original stimulus for such a program came from evaluations, beginning with the 1969 Westinghouse study, that were interpreted as showing that children soon lost the gains they made in Head Start because their later school experience did not maintain and build upon their earlier experience.

At any single site, PDC involves collaboration between at least one Head Start center and one public elementary school. The program emphasizes delivery of a variety of services that all converge on enhancing children's development. The services include the school curriculum itself, health care, care for the special needs of individual children, parent training, and staff training. It is an important aspect of PDC that these services are supposed to be organized on a broad cooperative basis, involving teachers and aides from all grades, Head Start through third, Head Start and school administrators, parents, and members of the general community. This merging of efforts is the mechanism considered most necessary to creating continuity of experience for children.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

[2]The work summarized here was performed pursuant to Contract No. HEW-105-75-1114 with the Administration for Children, Youth, and Families, Department of Health, Education, and Welfare.

Description of the Evaluation

ACYF's vision of the evaluation called for investigation of:

- the process by which the program was implemented,
- the nature of the resulting program,
- the program's effects on children's social competence,
- the program's effects on parents and teachers, and
- connections between elements of the program and its measured effects.

This was all to be accomplished by means of a design in which children from each site's treatment group would be followed for five years--from Head Start through grade three--and compared at intervals with a parallel group of children from other Head Start centers where the PDC treatment was not being implemented. These latter children, designated the comparison group, were to be comparable to children in the PDC group in all important respects except that after Head Start they were to attend elementary schools not participating in PDC.

Under this design, the program's effects on children could be determined by comparing PDC with non-PDC children; its effects on parents by comparing parents at PDC schools with parents at non-PDC schools; and its effects on teachers in a like manner. The Implementation Study, the part of the evaluation focusing on the process and result of program implementation, was to be mainly descriptive, not involving formal comparison of PDC schools with non-PDC schools.

From the beginning, ACYF had the worldly wisdom to recognize that life is not always as neat as evaluation plans would have it, so the first commission of the evaluation was not to evaluate the program, but to evaluate the evaluation. That is, the main task under our contract with ACYF was to determine whether or not it was feasible to do a program evaluation at all.

Ordinarily, evaluation is a pressing matter, and we who carry it out don't have the luxury of being able to ask whether we <u>should</u> or not. But the PDC program was designed so that the sites had a two-year warm-up period to prepare their programs, then a third year to make an actual run. The evaluation had the same advantage. It wasn't until the third and last year of our contract that the target group of children entered Head Start. At that point, the question was not so much how the <u>program</u> was working, but whether the <u>evaluation</u> would work if it were to follow the children for five years, until they completed the third grade. Our reports over the three-year term focused mainly on this issue and on related issues having to do with measuring PDC's effects on parents and teachers and with carrying out the study of program implementation.

Impact of the Evaluation

There have been a gratifying number of instances in which the PDC evaluation has had a positive impact on the decisions made about PDC, and some other instances in which it has apparently had little or no impact. Our experience over the past three years of the project suggests that four particular factors affect the impact of evaluation:

- the implications of objective evidence--meaning dependable, quantitative data relevant to decision criteria established beforehand,
- the social and political acceptability of what the evidence implies,
- the compatibility of the findings with what people believe in the first place, and
- the relevance of the findings to issues that have escaped formal definition--issues that are important but have not been expressed in terms of decision criteria.

Pure objective evidence is often spoken of as the only factor that matters, or ought to matter, in decision making. But this probably represents an unreasonably narrow view of what it means to make sensible decisions in a complex social environment. The history of the PDC evaluation offers some insight into the complexities of decision making.

Instances of Impact and of the Absence of Impact

ACYF's decision to continue the evaluation. The conclusion we eventually came to in our reports was that it was indeed feasible to conduct a long-term evaluation of PDC--that it would produce plausible and useful results. The conclusion was based mainly on our study of the tests that would be used, on assessments of the number of children available to PDC and their likely attrition, and on the suitability of the research design, which involved finding a comparison group of children in each site who were enough like the PDC children to be considered essentially the same. ACYF found this evidence convincing also, and accepted the recommendation that the evaluation be continued. Sometime this year a contract will be awarded for the work.

Adaptation of the Implementation Study's findings as formative feedback. Another instance of positive impact is the unexpected adaptation of findings from the PDC Implementation Study as a basis for program modification. This was unexpected because the rationale for documenting program implementation had been, first, to assess the character of the treatment in order to clarify causal connections between treatment and

effects, and, second, to document the program's history in a way that would allow lessons to be drawn from it for the sake of future programs. But instead, our descriptions of PDC's progress and our tentative analyses of the underlying dynamics have been converted to use as immediate formative evaluation data by the program's local and national staff. Although the descriptions and analyses have by no means passed without challenge, we are told that they have stimulated a number of self-improvement efforts. In April, personnel from all the PDC sites will be meeting in a national conference organized for purposes of refining the program, and the first item on the agenda is to be discussion of the Implementation Study's implications. The Implementation Study was never meant to have this effect this soon--it wasn't in the contract--but it seems to have answered an unanticipated need.

Failure to influence OMB. One place where we haven't had an impact, though, is on the Office of Management and Budget (OMB). This is the agency that approves the collection of data under federal programs--or in many cases disapproves it. When we applied for OMB clearance to collect Implementation Study data and survey data at the PDC sites, we provided that agency with the same evidence of the study's feasibility that we had provided to ACYF. For example, we presented analyses showing the characteristics of the treatment and comparison groups side by side on 16 variables such as ethnicity, prior school experience, and baseline performance on a variety of tests. On all but one or two of these variables the groups were statistically identical or at least very similar. But whereas ACYF concluded that this was evidence of the overall similarity of the groups, OMB thought it was evidence of some essential difference. Because of this and other differences in interpretation, OMB refused to approve several important parts of our data collection plans. Fortunately, we were able to accomplish the main purpose of the evaluation in ways that weren't prohibited by OMB.

Failure to interest some federal officials who have jurisdiction over PDC. Despite the evaluation's formative influence on staff directly responsible for the operation of the demonstration project, the reports have received little attention from federal officials at other levels of decision making. Since the issues addressed by the evaluation were defined originally by determining what federal officials needed to know to make decisions, it is natural to ask why, when the findings came in, they didn't seem to tell some officials anything they wanted to know. Because our experience leaves no doubt that all of those connected with PDC have sincere concerns for the program's welfare, we must conclude that some of these concerns have escaped the evaluation's attention. But since the domain of attention was defined by the decision makers in the first place, this conclusion has important implications for the design of effective evaluations. Following is a discussion of what our various experiences imply for the conduct of evaluation.

Analysis of the Dynamics of Impact

Pertinence to preset decision criteria. First, we carried out the evaluation in a way that provided objective information of high quality--in our own judgment and in the judgment of the outside consultants who made up our advisory panel. In doing this we were observing the usual rational tradition of evaluation. But however important this may be to producing evidence that others will consider objective, it doesn't seem to be enough by itself. Objective evidence is one factor affecting an evaluation's impact, but only one.

Social and political acceptability of the evaluation's implications. Another is the social and political acceptability of what the evidence implies. It has to persuade not just the people who ostensibly make the decisions, but also the people they have to persuade. The support our findings got from our outside advisory panel helped to establish the findings as credible to a larger audience than ACYF itself, and this gave the agency some assurance that what they found believable would be believed and accepted by the others who mattered. To restate this as a principle: evaluators should recognize that decisions are made in a social and political context, and no matter how convincing the evidence appears to the decision makers, decisions are constrained by what others believe and will allow. One way we can take account of this is to build into our evaluations ways of representing these social and political currents. Establishing a panel of recognized experts is one way. Keeping in touch with the first-line decision makers is another.

Congruence with predispositions of the audience. Another factor that affects an evaluation's impact on its audience is what the audience believed in the first place. The closer the evaluation findings are to the audience's expectations, the easier it is for the audience to believe in them. And conversely, the farther they are, the harder it is for the audience to reconcile them with what they expected.

ACYF expected that it would be possible to conduct a long-term evaluation of PDC. It wouldn't have been justifiable as an experimental program otherwise. So our conclusion that a solid, useful evaluation could be done was quite credible to them. But OMB's job is to keep federal data collection to a minimum, and they didn't accept what seemed to be the clear implications of our findings, possibly because of an institutional disposition not to.

One lesson in this may be that evaluators need to be attentive to the moods of their audiences. Evidence is in the eye of the beholder, in other words, and the greater the discrepancy between what the beholder expects and what we, as evaluators, think our findings imply, the greater the burden on us to present our findings in a way that does the most to reconcile the differences.

Considering how complicated life is in any organization, it seems unlikely that a formal evaluation can really address all the subtle issues that turn out to affect organizational decisions. But one way to at least make an evaluation as sensitive as possible to the variety of influences that affect decisions is to interact often with the people who will have to make the decisions. This allows for continuing review of what the evaluation is accomplishing and occasional revision of what it aims to accomplish.

Pertinence to latent decision criteria. The fact that the Implementation Study has become influential when it wasn't intended to be until after the program ended suggests another principle that deserves attention: when the focus of an evaluation is predetermined, as it is if a strict objectives-based model is followed, some critical decision-making criteria may be overlooked, either because they didn't exist at the outset or because they simply weren't expressed then. By the same token, an evaluation may incidentally attend to factors that were not emphasized at the outset but that eventually do assume significance. Thus although the Implementation Study was designed to satisfy long-term program design purposes, it has turned out to be of great interest to PDC program staff because it happens to fulfull an unanticipated need for guidance of current program operations. On the other hand, neither this emphasis nor any of the other predetermined evaluation emphases seem to have been sufficient to arouse interest in our findings at other levels of the federal hierarchy. This, presumably, is because the issues now considered critical at those levels were overlooked when the evaluation was designed, or have emerged since then.

Summary

To summarize, I'd say that there are four separate factors that have affected the impact of the PDC evaluation so far. First, the bearing of quantitative data on predetermined decision criteria. The conventional task of research or evaluation is to produce factual evidence on matters that are officially critical. The PDC evaluation succeeded in producing this kind of evidence, which is objective by definition, and was therefore conducive to objective decisions.

But besides serving decision making directly, this evidence also contributed to another factor that affected impact, the social and political acceptability of the evaluation's implications. Program decisions are made in a social and political environment, and those who make the decisions have to contend with the forces that operate in that environment. The PDC Advisory Panel's endorsement of the evidence helped to strengthen the continuation decision in this respect. To state the case another way: an evaluation must not only persuade the decision maker, it must also help to persuade the people the decision maker has to persuade. So far, as I said, it has not succeeded in persuading the Office of Management and Budget, which is certainly a force to contend with.

This difficulty brings up a third factor affecting the evaluation's impact, congruence with the predisposition of the audience. It's easier to convince someone of something that is congruent with a prior belief than it is to convince them of something contradictory. OMB, for its own reasons, seemed unfavorably predisposed toward PDC. ACYF, on the other hand, was inherently favorable toward its own program. And these two agencies, with opposite dispositions, came to opposite conclusions on the basis of the same evidence. This suggests a rule: the easier it is for the audience to reconcile its earlier beliefs with what the evaluation implies, the more believable those implications will be.

One last factor of impact that I've pointed to is pertinence to latent decision criteria. By this I mean the evaluation's bearing on matters that are of importance to the audience but that are not officially recognized as important. Sometimes they're not recognized because they have emerged late in the game, and sometimes it's because they're not supposed to be recognized too openly. For example, common sense will affirm that any government agency's programs are expected not only to accomplish their official objectives, but also to make a good impression on those outsiders who control the agency's future. If the PDC evaluation has not fascinated some federal officials as we'd hoped, it may be because our focus on formal decision criteria has caused us to overlook some unstated issues that are actually important to these officials. I've already given a contrasting example of what happens when an evaluation does, by chance, strike an important but unidentified interest of the audience: the Implementation Study was spontaneously converted by PDC staff from a historical record to an immediate formative evaluation of local PDC programs.

Epilog

These are some of the lessons about evaluation's impact that I would pass on from the PDC evaluation. Some of the lessons were learned by honoring the principles described here, some were learned, regretfully, after failing to honor the principles described. In the end, we have to recognize that evaluations of social programs can't hope to make all the difference in what is eventually decided about these programs, but we could make more of a difference if we were more observant of what it is that moves others--and ourselves--to decisions of consequence.

The Impact of the Evaluation of Project Follow Through Cognitively Oriented Curriculum[1,2]

Allen G. Smith
High/Scope Educational Research Foundation
Ypsilanti, Michigan

Each of the evaluations represented on this panel has occurred either within the context of a third-party national evaluation contracted by a federal agency, or as a part of smaller scale in-house research on demonstration programs run in Ypsilanti by High/Scope. Further, all of the evaluations are of programs of relatively short duration; the Perry research has continued for years, but the program itself was relatively brief. In contrast, High/Scope's Follow Through sponsor evaluation has occupied a peculiar position half-way between small-scale research on our own program, and large-scale evaluation of someone else's. Further, the program itself has continued for ten years; during this time the curriculum, the evaluation, and the audience have changed considerably. Consequently, there have been many threads to the evaluation through the years, and many real or potential audiences. That the evaluation has had any impacts at all is primarily the result of a reasonably tight design and, more importantly, of the symbiotic relationship that was created in-house between Follow Through research and curriculum development. The purpose of this paper is to explore the effects of the evaluation on these diverse audiences, and to reflect on the reasons for whatever impacts or lack of impact the evaluation has had.

The National Follow Through Project

The national Follow Through Project is the largest educational experiment ever undertaken. Since it began in 1967, approximately 500 million dollars has been devoted to implementing Follow Through projects in over 150 school districts in every state in the Union. Follow Through was initially intended to be a broad-based service program similar to Head Start. However, when the expected first-year appropriation of $120 million was cut to $15 million, the project was reconceptualized as a "planned variation experiment" comparing systematically varied curricula.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

[2]The work summarized here was performed pursuant to Grant No. G00750225 with the Office of Education, Department of Health, Education and Welfare by the High/Scope Foundation.

To accomplish this planned variation, independent curriculum developers (including some of those who participated in "Planned Variation Head Start") were brought in to "sponsor" particular educational models. Grants were made to each educational sponsor to quickly develop an innovative kindergarten through third grade program and to work with selected local projects to implement the sponsor's curriculum model. Altogether, there have been 23 sponsors, each affiliated with several local projects (some local projects were not associated with any sponsor).

As a Follow Through sponsor, High/Scope has worked with a total of ten local projects through the years (currently seven) to implement the early elementary version of the Cognitively Oriented Curriculum--an open framework (more commonly, "open classroom") approach strongly influenced by the theory of Jean Piaget. High/Scope's ten projects vary considerably, ranging from large northern, inner-city schools to rural schools in the south.

Evaluation in Follow Through

Because it was conceived as a planned variation experiment, the commitment to evaluation in Follow Through has been as massive as the program itself. Federally supported evaluation of the program was to have been conducted by a national evaluation contractor and by the individual sponsors. The national evaluation of Follow Through has had a long and somewhat controversial history, and will be discussed here only fleetingly. The original contractor was SRI International; in 1972 responsibility for data analysis shifted to Abt Associates, Inc. The important feature of the national evaluation for our purposes has been what Haney (1977) called the "funnelling effect." From an early commitment to a broad range of variables, the national evaluation has steadily narrowed its focus until in Abt's final report to USOE, for example, judgments of sponsor impacts on children's academic achievement were based on only one measure, the Metropolitan Achievement Test, examined within a single analytic design.

From the beginning, Follow Through sponsors were permitted to conduct evaluations of their own Follow Through programs. As it became apparent, however, that the national evaluation would not be able to evaluate each program "in its own terms," USOE began to encourage sponsors to expand their evaluation efforts. By 1976, 25% of High/Scope's total Follow Through budget was devoted to evaluation.

High/Scope's evaluation efforts have had three main thrusts:

- evaluation of children's performance on traditional measures of aptitude and achievement;
- assessment of classroom process variables through implementation checklists and time sampling observation systems;
- development of alternative measures of children's ability to express their thoughts and experiences in writing.

Initially, High/Scope's own evaluation relied almost exclusively upon existing measures of academic aptitude and achievement and, therefore, closely resembled national evaluation efforts in its scope. Gradually, however, as measures of educational process and measures of alternative child outcomes were developed, the focus shifted away from traditional measures. Today, the only standardized aptitude and achievement test data analyzed by High/Scope are those collected routinely by the local school districts.

Several research designs have been employed through the years. All have been quasi-experimental, given the impossibility of randomly assigning children to Follow Through and non-Follow Through treatment conditions. In some instances, Follow Through children have been compared with children attending the same school prior to the implementation of the Follow Through program. In other cases, Follow Through children have been compared with contemporary non-Follow Through children matched on demographic and economic variables. Yet a third design involved comparisons of successive cohorts of Follow Through children to ascertain the effects of changes in curriculum model implementation.

The audience for sponsor evaluations has never been explicit. Initially, and at various times since, it was thought to be the national Follow Through evaluation staff at USOE. At other times, other audiences have been considered:

- The field of early childhood education
- Local school boards
- The Joint Dissemination Review Panel (JDRP)
- High/Scope curriculum developers

No one audience was ever specified by the funding agency as paramount; sponsors have been free to choose their own.

Impacts of the Evaluation

In ten years, High/Scope's Follow Through sponsor evaluation has had a variety of impacts on its several audiences. Some of these were anticipated; others were not. Also, the evaluation has had discouragingly little impact in several key areas. Through hindsight, some of the reasons for these successes and failures can be surmised, but before turning to the reasons, let us review quickly some of the impacts.

Impacts at the Level of National Follow Through

Sponsor evaluations could potentially have had important impact on the national evaluation of Follow Through. Since the national evaluation was so restricted in its focus, sponsor collected data could have been used to provide a more balanced picture of the effects of the program. High/Scope's evaluations have had some effects on interpretations made by USOE's Office of Planning, Budget, and Evaluation (OPBE), the agency that since 1974 has been responsible for Follow Through evaluation. High/Scope, along with other sponsors, was invited to prepare a chapter for a volume of USOE's final evaluation report, and some of these data will be included in the USOE synthesis of evaluation findings. The weight of sponsor findings caused the authors of this volume to judge the overall findings (national and sponsor) for High/Scope "inconclusive," rather than generally ineffective as suggested by the national evaluation data alone. Unfortunately, the Abt final report of the national evaluation was released to the media long before the rest of the USOE report (which has still not been released). Consequently, public and congressional impressions of the effectiveness of the program have been shaped without the benefit of sponsor data.

Another effect of sponsor evaluations at the national Follow Through level has come, not from their data, but from the fact that the sponsors had research staffs competent to review and provide technical critiques of the national evaluation. This is uncommon in national evaluations. The effects of this give and take are sometimes hard to see, but through the years the sponsor research staffs have had some hand in shaping the national evaluation designs and reports. Sponsors have provided some detailed technical critiques of draft evaluation reports; their reactions have on several occasions prompted reanalyses by Abt. Most recently, representatives from the sponsor group have prepared a document that reanalyzes certain of the national evaluation data and suggests some alternative conclusions. While it would be a gross overstatement to say that sponsors are satisfied with the final shape of the national evaluation, their input and critiques have helped make it less objectionable than it might otherwise have been.

Impacts on Local Sites

Local Follow Through projects could potentially have two uses for evaluation data. First, they could use positive findings of program impact to demonstrate their worth to district, federal, or other decision makers. Second, sponsor evaluation data could have been used formatively by the projects to improve their own programs.

In fact, data collected or analyzed by High/Scope have been used for both of these two purposes. Virtually all of our sites have requested at some time summaries of our findings for presentation to local district officials. Last year, data collected and analyzed by High/Scope were instrumental in securing validation for one project from the Joint

Dissemination Review Panel in Washington. As for formative impacts, there have been several. Presentation of the results from assessments of levels of implementation have at several sites prompted training for teachers in areas that the assessment found lacking. Findings at some sites of no difference between Follow Through and non-Follow Through children have provoked different responses. At a few, principals and project directors have responded with new initiatives aimed at increasing levels of program implementation. At others, the reverse has occurred: principals have taken the findings as evidence that the program does not work and ordered a reversion to more didactic instructional approaches. In either case, the evaluation results did stimulate changes in the local projects, although some were not what we would have wanted.

Impacts on High/Scope

Effects of the ten years of sponsor research on High/Scope itself can be seen on curriculum development and on High/Scope's approach to education research in general.

Effects on curriculum development have been different from what one might expect in a research and development organization. We like to think of curricula being developed based on the findings from research, but in this case the curriculum was affected more by the efforts that went into developing instruments sensitive to program objectives than by the data those instruments produced. The reason for this lies in the history of High/Scope's participation in Follow Through. Like most sponsors, High/Scope did not have a fully-developed curriculum model when Follow Through began. Meanwhile, sponsor-based research was supposed to be measuring whether the curriculum approach was having the desired effects on children. For High/Scope, this meant that traditional tests of academic achievement could not be the only child outcome measure. To make impact findings interpretable, moreover, measures of degrees of classroom implementation had also to be developed. Development of all of these instruments demanded close collaboration between research and curriculum staffs within the Foundation. Since the curriculum was being developed at the same time, these meetings proved as beneficial for the program staff as they were to research. From them came more systematic statements of the goals and methods of the Cognitively Oriented Curriculum.

The effects of High/Scope's sponsor evaluation on research at the Foundation have been more apparent than its effects on the curriculum. Follow Through and educational evaluation have grown up together, both nationally and within High/Scope. Many lessons have been learned from ten years of sponsor evaluation about how one goes about researching alternative approaches to education and about the problems and pitfalls of quasi-experimental research. These lessons have been applied to other work within the Foundation. The BEH First Chance Project used an adaptation of the Productive Language Assessment Tasks. The implementation

assessment instrument originally developed for Follow Through was subsequently adapted for use by the Foundation's preschool department. Returning to the PLAT, High/Scope now has a grant from the Carnegie Corporation to pursue further the development of measures of children's productive language.

Effects in the area of evaluation design and analytical methods are harder to trace, but it would be safe to say that certainly the two national evaluations conducted by High/Scope--PDC and Home Start--have benefited greatly from the Follow Through experience.

Impacts on the Field of Early Childhood Education

Unlike several of the studies represented here, High/Scope's Follow Through sponsor evaluation has had minimal impact on the world outside Follow Through. Although the language assessment instruments developed for Follow Through have received considerable attention outside the Foundation, and have been used by others, for the most part the results of the research have not been widely disseminated.

Reasons for the Impact or Lack of Impact

Through ten years, the High/Scope sponsor evaluation has had a variety of impacts on several audiences, although not as much on some as we would have liked. In retrospect, several reasons for this mixed performance can be seen. First, the evaluation was conducted using research designs that were as rigorous as the situation allowed, and these research designs did produce findings on program effect. Consequently, we were able to help some sites demonstrate their effectiveness to school boards and to JDRP. Because of problems with implementation and adequacy of comparison groups at some sites, though, we were not able to truly test the curriculum elsewhere. When the program was implemented, and where there were adequate comparisons, Follow Through favoring effects were found, but this happened at fewer than half our sites. Nonetheless, the findings were sufficient to raise serious doubts about the null effects reported by the national evaluation.

Second, the audience for sponsor research has never been clear. Evaluations performed under contract to an agency, like Home Start and PDC, by their very nature have a defined audience. Even some grant-supported research, like the Perry longitudinal Study, has little ambiguity about audience. Follow Through sponsor research, by virtue of its peculiar position somewhere between a national evaluation and in-house research, had many potential audiences, but none was specified. At various times the audience was thought to be USOE decision makers, at others it was thought to be the "field," and at still others it was thought to be local sites or the Joint Dissemination Review Panel. Each of these groups has different needs, and by meeting the needs of one the findings ran the risk of becoming irrelevant to others. By having research that partially met the needs of all, the needs of none were met squarely.

Finally, many of the effects described here were made possible by the symbiotic relationship that existed between High/Scope's evaluation and curriculum development. Close administrative and proxemic relationships between research and program staffs have insured that the evaluation was sensitive to the objectives and methods of the curriculum; this collaboration has resulted in development of compatible instruments and in a mutually reinforcing relationship between program and instrument development. Too often in education, the evaluation "tail" is permitted to wag the programmatic "dog." We measure what we can measure rather than what should be measured. By housing evaluation with curriculum developers at High/Scope, evaluation technology and the Cognitively Oriented Curriculum were permitted to evolve together. The presence of researchers also established an advocacy group at the sponsor level who could respond substantially to issues and problems in the national evaluation.

What are the lessons from this evaluation in general? Probably the two most obvious lessons echo those of other presentations on the panel: evaluations are most effective when they satisfy certain minimal design criteria and when the audience for them can be clearly identified and their information needs anticipated. But there is another lesson that is perhaps not so obvious: close relationships between evaluation research and curriculum can be healthy ones, and their benefits outweigh the dangers of bias. Such relationships help guard against evaluations that completely miss their mark in selecting variables for measurement, as did the national Follow Through evaluation of certain models; such relationships also benefit curriculum development, because the definition of variables for measurement is necessarily a process of clarification and reflection. Finally, such relationships speed the development of evaluation instruments sensitive to the processes and objectives of alternative education programs.

The Impact of the Evaluation of the Ypsilanti Perry Preschool Project[1]

Lawrence J. Schweinhart
High/Scope Educational Research Foundation
Ypsilanti, Michigan

As I reviewed the papers of my colleagues who have been involved in national evaluations, it became clear to me that we have a striking difference in perspective about audience. They are trying to influence specific decisions of specific policy-makers within the federal government. The Ypsilanti Perry Preschool Project today is not directed at such tightly defined targets. In its early days, it was directed at specific local and state school personnel. But today its audience is the field of early childhood education. It has a broad audience, both national and international. It is frequently referenced in scholarly journals, undergraduate textbooks, and even the mass media. This paper describes some of the broad impacts of the Perry Project evaluation and some of the reasons behind its impact.

The Perry Project Evaluation

The Ypsilanti Perry Preschool Project operated from 1962 through 1967. During that period, five waves or cohorts of children cycled through the preschool program for two years apiece, so that a total experimental group of 58 children attended preschool. During the same period, a control group of 65 children in five waves was accumulated; they received the same assessments as the experimental group, but did not attend preschool. Both groups were selected to have a low tested IQ at age 3 and to have come from families of low socioeconomic status.

Children were assigned to groups on the basis of their IQ scores with adjustments to equate the sex ratio and average socioeconomic status of the groups. The technique proved highly successful in that groups did not initially differ on IQ, socioeconomic status, family size, mother's education, or a number of other important demographic variables. (Maternal employment did differ across the groups, but the difference was found to diminish, rather than increase, apparent program effects.)

The preschool was in operation five half-days a week, with weekly home visits to the families involved. The preschool program emphasized individualized support of each child's cognitive development, an approach which became the basis of High/Scope's Cognitively Oriented Curriculum. Staff met regularly to discuss program operation and to engage in problem-solving as a group.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

The longitudinal evaluation of the project is now in its sixteenth consecutive year, a rare accomplishment in longitudinal research. The duration of the Perry Project has paradoxically enabled it to be a trend-setter on numerous occasions. It was one of the first projects to indicate that preschool could raise IQ (Weikart, Deloria, Lawser, & Wiegerink, 1970); thereby it contributed to the early enthusiasm for Project Head Start. It was one of the first projects to indicate that preschool IQ gains would not last forever. Recognition of this fact led not only to a pessimism about preschool, but also to a general pessimism about all compensatory education programs and renewed arguments about a genetic component of intellectual competence.

Now the Perry Preschool Project may well be on the crest of a new wave, a mood of cautious optimism about the long-term effects of preschool programs for disadvantaged children. It is becoming clear that the Perry Project demonstrates a lasting scholastic and economic impact of preschool. By fourth grade, preschool experience reduced by half the number of children retained in grade or placed in special education--38% of the control group vs. 17% of the experimental group. The financial cost of operating the Perry Preschool was more than offset by the financial benefits to society of that preschool experience (Weber, Foster, & Weikart, 1978). The benefits entered into the analysis were: reduced cost of special education and grade retention; released time of mothers during preschool; and increased lifetime earning potential predicted from the increase in scholastic success. Preschool experience also accounted for significant achievement test gains throughout school; our last assessment at eighth grade indicated a difference of 1.2 grade equivalents favoring children who attended preschool (Schweinhart & Weikart, 1977; Weikart, Bond, & McNeil, 1978).

Other researchers were also contributing to this new spirit of cautious optimism through the long-term evaluations of their preschool projects. In the Consortium of Longitudinal Studies (1977), a group to which we belong, similar favorable findings on reduced grade retention and special-education placement have been found in the projects of Ira Gordon, Phyllis Levenstein, Susan Gray, Frank Palmer, and Edward Zigler.

Reasons for Evaluation Impact

The Perry Project has been influential in forming a climate of support for preschool education. Three reasons for this influence may be identified:

- The evaluation was an intensive study, designed with particular concern for internal validity.
- The program had strong effects.
- The evaluation was timely and long-lasting.

First, the evaluation was an intensive study, designed with particular concern for internal validity. Internal validity refers to those aspects of a design which make findings interpretable and give us confidence that the findings are accurate. The key feature of the Perry Project's internal validity is the use of an experimental group and a control group that were highly similar on major background characteristics. The project had a replication element--five waves of children cycling through the classroom over the years--which strengthened its generalizability. The replication element permitted the generation of a sufficient sample size at a single site. Because the program operated at only one site, it could be carefully monitored so as to insure a strong and consistent program implementation.

The Perry Preschool was not intended to be representative of some larger group of existing preschools, but rather was intended to serve as a model of what other preschools could become. The project does not address the question; Does preschool work all the time? Rather it addresses the question; Can preschool work under certain conditions? It was not until years later that people discovered the problems of implementing programs at a number of sites at the same time. Nonetheless, the Perry Project provides clear evidence that successful programs can be implemented. And, sometimes, multi-site studies buy greater generalizability at a cost which is too great: problems of implementation and design swamp their results to such an extent that firm conclusions are impossible.

Second, the program had strong effects. Program effects have been found on some important measure consistently since the first year of the program. Effects have been found across a range of measurement areas. It is important to find areas of measurement where programs are effective and to document successful educational programs. In this way educational research is in a position to advocate clear directions to teachers and educators.

Third, the evaluation was timely and long-lasting. The major factor here may be that the evaluation was relatively independent of political urgencies. Over its history, the evaluation has received external funding from the Office of Education, the Spencer Foundation, the Carnegie Corporation, and now the Administration for Children, Youth, and Families. Funding was given in approval of our evaluation plans, not as leverage for external control. There were times when the evaluation did not receive external funding. With or without funding, the evaluation design has been maintained and data collected on schedule.

Recommendations for Future Evaluations

The reasons for the impact of the Perry Project evaluation translate directly into recommendations for future evaluations of educational programs.

First, intensive one-site studies with high internal validity can be effective and ought to be supported. In the Perry Project, the balanced experimental/control group design permits firm inferences about program effects 14 years and more after children completed the program. The replication element in the design permitted us to generate a sizable sample of 123 children while closely monitoring a single site. One-site studies with annual replications may well be the design of choice when one is trying to implement and evaluate a complex, subtle curriculum.

Second, before programs are committed to longitudinal evaluation, they should be scrutinized for their likelihood of producing lasting impacts. Such scrutiny might be carried out by a review process similar to that of the Joint Dissemination and Review Panel. The review process ought to involve onsite observation, interviews with staff, and examination of early outcomes of the project. Approval for longitudinal status should probably not be given until the program to be evaluated has been in operation for several years. In this way, initial implementation problems could be solved, and short-term evaluations could provide data either for or against the likelihood of long-term impact.

Third, funding agencies should set aside funds for longitudinal evaluative research which are stable and resistant to political urgencies. Presently it requires rare determination to try to carry out longitudinal research. Yet everyone acknowledges its worth. Without longitudinal research, we would be left with the hastily formed conclusion that a good preschool has only a short-term effect on IQ, rather than the powerful lasting effects we are now discovering. Longitudinal research would unquestionably be strengthened by funding patterns which encourage continuity of staff and fidelity to a long-term research design.

With the exception of a handful of truly longitudinal efforts, much evaluative research in early childhood education over the past decade has been in a great hurry. Unfortunately, the hurry may not have added as much as we had hoped to our substantive knowledge about what works in early childhood education. Inasmuch as the Perry Project has made a contribution, it has also taught us that patience is a most useful virtue in evaluative research.

References

Consortium of Longitudinal Studies. The persistence of preschool effects. Washington, D.C.: Administration for Children, Youth, and Families, 1977. (Final report; Grant No. 18-76-07843).

Schweinhart, L. J., and Weikart, D. P. Research report: Can preschool education make a lasting difference? Bulletin of the High/Scope Foundation, No. 4, 1977.

Weber, C. U., Foster, P. W., and Weikart, D. P. An economic analysis of the Ypsilanti Perry Preschool Project. Monographs of the High/Scope Educational Research Foundation, No. 5, 1978.

Weikart, D. P., Bond, J. T., and McNeil, J. T. The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade. Monographs of the High/Scope Educational Research Foundation, No. 3, 1978.

Weikart, D. P., Deloria, D. J., Lawser, S. A., and Wiegerink, R. Longitudinal results of the Perry Preschool Project. Monographs of the High/Scope Educational Research Foundation, No. 1, 1970.